Phân Tích Dữ Liệu Thực Tế với Python

Bài 7.2: Xử Lý Dữ Liệu Nâng Cao với Pandas

Quang-Khai Tran, Ph.D
CyberLab, 03/2023

(Ảnh: Internet)

CYBERLEARN
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

CYBERSOFT
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

CYBERLAB
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

Nội dung

1. Dữ liệu không phải dạng số
2. Một số xử lý thống kê
3. Làm sạch dữ liệu
4. Bài tập & Thảo Luận

Phần 1.
Xử lý một số dữ liệu không phải dạng số

1.1. **Dữ liệu text**
1.2. **Dữ liệu datetime**
1.3. **Dữ liệu dạng phân loại** (categorical)

1.1 Dữ liệu text

Pandas hỗ trợ 2 cách để lưu trữ dữ liệu dạng text trong bảng:

- Định dạng `object` của mảng Numpy
- Định dạng `StringDtype` mở rộng

- Tham khảo thêm:
  https://pandas.pydata.org/pandas-docs/stable/user_guide/text.html

1.2 Dữ liệu datetime

- Thường sử dụng datetime như là index của dữ liệu time series
  - Có thể chọn các dòng theo thời gian 1 cách thuận tiện

1.3 Dữ liệu dạng phân loại

Là dữ liệu dành cho một mảng giới hạn số lượng các giá trị dạng string:

- Ví dụ: ['good', 'bad', 'very-good']
- Thường dùng trong một số trường hợp đặc biệt, chẳng hạn như báo hiệu cột tương ứng là dạng categorical cho các thư viện khác

- Tham khảo thêm: https://pandas.pydata.org/pandas-docs/stable/user_guide/categorical.html

Phần 2.
Một số xử lý cho thống kê

2.1 Ghép nối/trộn các dataframe

pd.concat	Ghép dataframe theo chiều dọc hoặc ngang
df.append	Ghép 1 dataframe vào 1 dataframe khác theo chiều dọc
df.join	Ghép 1 dataframe vào 1 dataframe khác theo chiều ngang
pd.merge	Ghép 2 dataframe dựa trên 1 cột có thông tin chung

2.2 Một số hàm thống kê thông dụng

Các hàm thống kê mô tả	`describe, min, max, mean, median, mode`
Thống kê sự tương quan	`corr(), corrwith()`
Tìm n sample (dòng) lớn nhất được sắp theo cột	`df.nlargest(n, columns)` (ngược lại: `df.nsmallest`)

2.3 Hàm value_counts

❖ Hàm value_counts()

```
df['tên-cột'].value_counts(normalize = False/True,
                           bins=None)
# df.column_name.value_counts().keys()/columns
# df.column_name.value_counts().index
```

2.4 Hàm groupby

Hàm groupby() thực hiện một trong các tác vụ sau:

- Chia/tách (splitting) bảng dữ liệu ra các nhóm khác nhau (theo chiều dọc)
- Áp dụng (applying) một hàm nào đó cho các nhóm một cách độc lập
- Kết hợp (combining) kết quả vào một cấu trúc dữ liệu

  ⇒ Thông thường nhất vẫn là chia/tách
- Thường kết hợp các hàm tính toán: sum, value_counts, mean, …

```
df.groupby([columns], sort=T/F, dropna=T/F)
```

Phần 3.
Làm sạch dữ liệu

3.1. Tìm và loại bỏ dữ liệu khuyết
- dropna

3.2. Điền vào dữ liệu khuyết
- fillna và ffill/bfill/backfill
- interpolate

3.3. Tìm và loại bỏ outliers

3.1 Tìm và loại bỏ dữ liệu khuyết

- Sử dụng hàm:
  isna(), isnull()
- Kết hợp các hàm sum, count để kiểm đếm số lượng phần tử/dòng thiếu

- Tham khảo nâng cao:
  https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html

3.1 Điền vào dữ liệu khuyết

- Sử dụng hàm:
  fillna(), interpolate()

3.3 Tìm và loại bỏ outliers

- Đối với dữ liệu theo phân bố chuẩn:
  Có thể loại bỏ các giá trị cách xa mean nhiều hơn 3 lần của Standard Deviation
- Đối với dữ liệu phân bố không đều:
  Tùy trường hợp, cần kiểm tra cẩn thận

4 Bài Tập

Trên dữ liệu NYC:

1. Bài 1:
   - Đếm số chuyến bay trễ của các sân bay (cột ‘origin’)
   - Tương tự, đếm số chuyến bay không trễ của các sân bay
   - Tạo bảng hiển thị số chuyến bay trễ vs. không trễ
   - Vẽ lên biểu đồ
2. Bài 2:
   - Thực hiện tương tự với các hãng hàng không (cột ‘carrier’)
3. Bài 3:
   - Tính thời gian trễ (gồm dep_delay và arr_delay) trung bình của các hãng
   - Tính thời gian trễ (gồm dep_delay và arr_delay) trung bình từ các sân bay
   - Vẽ các kết quả lên biểu đồ
4. Bài 4:
   - Tính tổng, trung bình, mean, median của quãng đường bay của từng hãng (cột distance)

THANK YOU!